Journées de Statistique. 44, 2012/05/21-25, Bruxelles, Belgique

Incertitude sur la segmentation pour les modèles de détection de ruptures multiples

Yann Guédon

CIRAD, UMR AGAP et INRIA, Virtual Plants, F-34398 Montpellier
E-mail: guedon@cirad.fr

Résumé. Le problème de détection *a posteriori* de ruptures multiples est étudié. En ce qui concerne l'incertitude sur la segmentation, les travaux se sont focalisés jusqu'à présent sur l'incertitude concernant la position des ruptures. Nous proposons de poser ce problème d'une façon différente en voyant les modèles de détection de ruptures multiples comme des modèles à structure latente et en utilisant des résultats de la théorie de l'information. Cela nous amène à montrer que les lois *a posteriori* de la position des ruptures ne reflètent que partiellement l'incertitude sur la segmentation. L'incertitude canonique est donnée par la loi *a posteriori* des segmentations. L'entropie correspondante peut être décomposée sous forme d'un profil d'entropies, ce qui permet de localiser cette incertitude canonique le long de la séquence. Nous proposons d'utiliser la divergence de Kullback-Leibler entre la loi uniforme et la loi *a posteriori* des segmentations pour des nombres de ruptures successifs comme critère de sélection du nombre de ruptures. L'approche proposée est illustrée par l'analyse des phases de croissance de pins laricio.

Mots-clés. Algorithm de lissage, Détection de ruptures multiples, Divergence de Kullback-Leibler, Entropie, Modèle à structure latente.

Abstract. The retrospective or off-line multiple change-point detection problem is addressed. Concerning the segmentation uncertainty in multiple change-point models, the focus was mainly on the change-point position uncertainty. We propose to state this problem in a new way, viewing multiple change-point models as latent structure models and using results from information theory. This led us to show that the posterior distributions of the change-point position only reflect partially the segmentation uncertainty. The canonical uncertainty is given by the posterior distribution of the segmentations. The corresponding segmentation entropy can be decomposed as an entropy profile which enables to localize this canonical uncertainty along the sequence. We propose to use the Kullback-Leibler divergence of the uniform distribution from the segmentation distribution for successive numbers of change points as a new adaptive criterion for selecting the number of change points. The proposed approach is illustrated by the analysis of Corsican pine growth phases.

Keywords. Entropy, Kullback-Leibler divergence, Latent structure model, Multiple change-point detection, Smoothing algorithm.

1 Introduction

Les modèles de détection de ruptures multiples sont ici vus comme des modèles à structure latente ce qui diffère du point du vue non-bayésien sur ces modèles où les ruptures sont vues comme des paramètres fixes à estimer. Une fois les ruptures estimées, il n'y a bien entendu plus de structure latente. Par rapport au point de vue bayésien sur ces modèles, l'approche proposée correspond au cas où toutes les segmentations pour un nombre de ruptures fixé sont supposées équiprobables, ce qui est une hypothèse classique (Liu et Lawrence, 1999; Girón *et al.*, 2007). Toute approche de détection *a posteriori* de ruptures multiples nécessite d'estimer les paramètres intra-segments pour les $T(T+1)/2$ segments possibles où T est la longueur de la séquence. Cette estimation peut être faite aussi bien dans un cadre non-bayésien que bayésien et on parlera alors de modèle hiérarchique empirique ou bayésien selon le paradigme d'estimation retenu. On s'intéresse ici à l'incertitude sur la structure latente, c'est à dire sur les segmentations possibles d'une séquence observée donnée pour un nombre de ruptures fixé, et le résultat principal est de montrer qu'on peut proposer une mesure d'incertitude canonique localisée le long de la séquence.

Deux types de processus d'états latents peuvent être définis pour un modèle de détection de ruptures multiples :

- Un processus à J états $\{S_t\}$ (J est le nombre de segments et $J-1$ le nombre de ruptures) représentant le rang du segment le long de la séquence (Chib, 1998) avec $P(S_0 = 0) = 1$, $P(S_{T-1} = J-1) = 1$ et $t = 1, \ldots, T-2 : P(S_t = j, S_{t-1} = i) > 0$ pour $j = i, i+1$,
- Un processus binaire $\{R_t\}$ où l'état 0 correspond à pas de rupture et l'état 1 à une rupture (Lavielle, 1998).

Un modèle de détection de ruptures multiples peut être vu comme un modèle à structure latente $\{S_t, X_t;\ t = 0, \ldots, T-1\}$ ou résumé par un modèle à structure latente $\{R_t, X_t;\ t = 0, \ldots, T-1\}$. Nous avons les relations suivantes entre les deux processus d'états latents :

$$\begin{aligned} P(R_t = 0) &= \sum_j P(S_t = j, S_{t-1} = j), \\ P(R_t = 1) &= \sum_j P(S_t = j, S_{t-1} = j-1). \end{aligned} \tag{1}$$

Dans la suite, $\mathbf{X} = \mathbf{x}$ désigne la séquence observée $X_{0:T-1} = x_{0:T-1}$ et $\mathbf{S}$ la segmentation $S_{0:T-1}$. Nous proposons d'utiliser l'entropie des segmentations $H(\mathbf{S}|\mathbf{X} = \mathbf{x}; J)$ comme mesure d'incertitude canonique sur le processus d'états latent.

2 Bornes sur l'entropie des segmentations et profils d'entropies

L'entropie des segmentations peut être décomposée en une somme de termes telle que chaque terme soit individuellement borné supérieurement

$$\begin{aligned} H\left(\mathbf{S}|\mathbf{X}=\mathbf{x};J\right) &= -\sum_{\mathbf{s}} P\left(\mathbf{S}=\mathbf{s}|\mathbf{X}=\mathbf{x}\right)\log P\left(\mathbf{S}=\mathbf{s}|\mathbf{X}=\mathbf{x}\right) \\ &= \sum_{t=1}^{T-1} H\left(S_t|S_{0:t-1},\mathbf{X}=\mathbf{x};J\right) \qquad (2) \\ &= \sum_{t=0}^{T-2} H\left(S_t|S_{t+1:T-1},\mathbf{X}=\mathbf{x};J\right). \qquad (3) \\ &\leq \sum_{t=1}^{T-1} H\left(R_t|\mathbf{X}=\mathbf{x};J\right) \end{aligned}$$

En effet, en appliquant des résultats classiques de la théorie de l'information (Cover et Thomas, 2006), nous obtenons les bornes supérieures suivantes sur les entropies conditionnelles,
$t=1,\ldots,T-1$:

$$H\left(S_t|S_{0:t-1},\mathbf{X}=\mathbf{x};J\right) \leq H\left(R_t|\mathbf{X}=\mathbf{x};J\right), \qquad (4)$$

$t=0,\ldots,T-2$:

$$H\left(S_t|S_{t+1:T-1},\mathbf{X}=\mathbf{x};J\right) \leq H\left(R_{t+1}|\mathbf{X}=\mathbf{x};J\right). \qquad (5)$$

Le décalage de 1 quand on conditionne par le futur est une conséquence directe de la définition des ruptures (1).

Pour chaque instant t possible, $H\left(R_t|\mathbf{X}=\mathbf{x};J\right)$ quantifie l'incertitude sur la loi *a posteriori* des ruptures $\{P\left(R_t=0|\mathbf{X}=\mathbf{x}\right), P\left(R_t=1|\mathbf{X}=\mathbf{x}\right)\}$ qui est un indicateur classique d'incertitude; cf. Fearnhead (2006). Les différentes entropies $H\left(S_t|S_{0:t-1},\mathbf{X}=\mathbf{x};J\right)$, $H\left(S_t|S_{t+1:T-1},\mathbf{X}=\mathbf{x};J\right)$ et $H\left(R_t|\mathbf{X}=\mathbf{x};J\right)$ peuvent être calculées par une extension de l'algorithme de lissage proposé par Guédon (2008). La complexité de cet algorithme est en $O(JT^2)$ en temps et en $O(JT)$ en espace. Les entropies $H\left(S_t|S_{0:t-1},\mathbf{X}=\mathbf{x};K\right)$, $H\left(S_t|S_{t+1:T-1},\mathbf{X}=\mathbf{x};K\right)$ et $H\left(R_t|\mathbf{X}=\mathbf{x};K\right)$ pour $t=0,\ldots,T-1$ et $K=2,\ldots,J-1$ segments peuvent être obtenues comme sous-produits de l'algorithme de lissage pour J segments et ceci sans changer la complexité de cet algorithme. Les décompositions (2) et (3) nous donnent deux moyens possibles pour décomposer l'entropie des segmentations sous forme de profils d'entropies conditionnelles. Le choix du conditionnement doit traduire des hypothèses sur la structuration des données. Les inégalités (4) et (5) montrent que les profils d'entropies conditionnelles $\{H\left(S_t|S_{0:t-1},\mathbf{X}=\mathbf{x};J\right);\ t=0,\ldots,T-1\}$ et

$\{H(S_t|S_{t+1:T-1}, \mathbf{X}=\mathbf{x}; J)\,;\ t=0,\ldots,T-1\}$ sont bornés point à point par le profil d'entropies des ruptures $\{H(R_t|\mathbf{X}=\mathbf{x}; J)\,;\ t=0,\ldots,T-1\}$. Comme $H(R_t|\mathbf{X}=\mathbf{x}; J)$ est borné supérieurement par $\log 2$, tous ces profils sont definies sur la même échelle absolue $[0, \log 2]$ ce qui les rend facilement interprétables. Enfin, l'entropie des ruptures $\sum_{t=1}^{T-1} H(R_t|\mathbf{X}=\mathbf{x}; J)$ est bornée supérieurement par $(T-1)\log(T-1) - (J-1)\log(J-1) - (T-J)\log(T-J)$ qui est une bonne approximation quand T est suffisamment grand de $\log n_J$ avec $n_J = \binom{T-1}{J-1}$, nombre de segmentations possibles pour J segments. Par contre, il n'y a pas de relation d'ordre fixe entre l'entropie des ruptures $\sum_{t=1}^{T-1} H(R_t|\mathbf{X}=\mathbf{x}; J)$ et l'entropie des segmentations sous une hypothèse de loi uniforme, $\log n_J$.

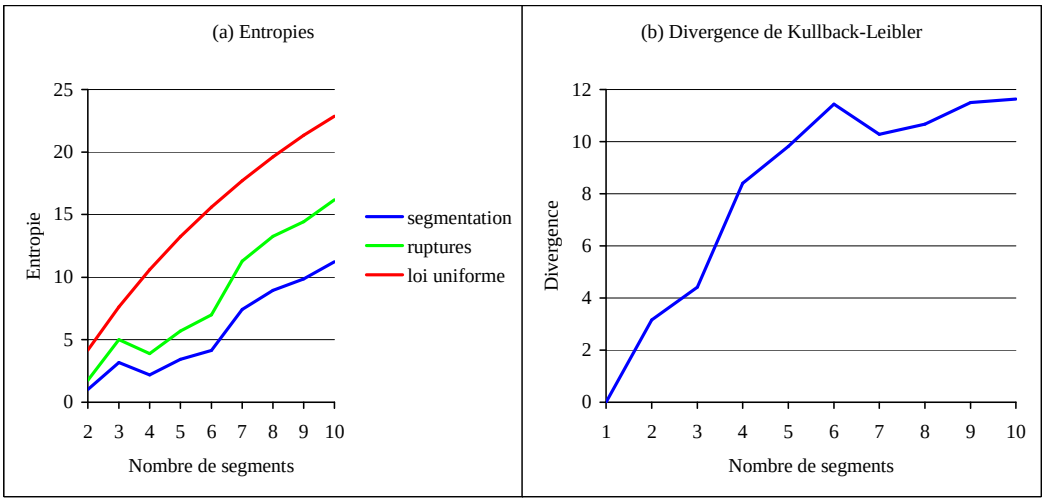

Figure 1. (a) Entropies et (b) Divergence de Kullback-Leibler fonction du nombre de segments.

3 Divergence de Kullback-Leibler pour sélectionner le nombre de ruptures

Dans l'esprit de la statistique Gap proposée par Tibshirani *et al.* (2001), nous proposons d'utiliser la divergence de Kullback-Leibler entre la loi uniforme et la loi *a posteriori* des segmentations pour sélectionner le nombre de ruptures

$$D_{KL}(P_J \,||\, U_J) = \log n_J - H(\mathbf{S}|\mathbf{X}=\mathbf{x}; J)\,.$$

Cette divergence peut être interprétée dans un cadre bayésien comme la divergence entre la loi *a priori* uniforme U_J et la loi *a posteriori* P_J. Ce critère empirique a émergé en constatant que, quand $J > J_{\text{optimal}}$, les segments étaient divisés de manière injustifiée vis à vis des données. Cela générait alors un très grand nombre de segmentations de probabilités *a posteriori* non-négligeables par rapport à la probabilité *a posteriori* de la

segmentation optimale en J segments. Par rapport aux critères de vraisemblance pénalisée dans un cadre non-bayésien et à la loi *a posteriori* du nombre de ruptures dans un cadre bayésien, la spécificité de ce critère est de reposer exclusivement sur l'incertitude sur la segmentation et d'être applicable à la fois dans un cadre non-bayésien et dans un cadre bayésien. Ceci ouvre en particulier des possibilités pour juger de la validité du nombre de ruptures obtenu par différentes méthodes de sélection de modèles.

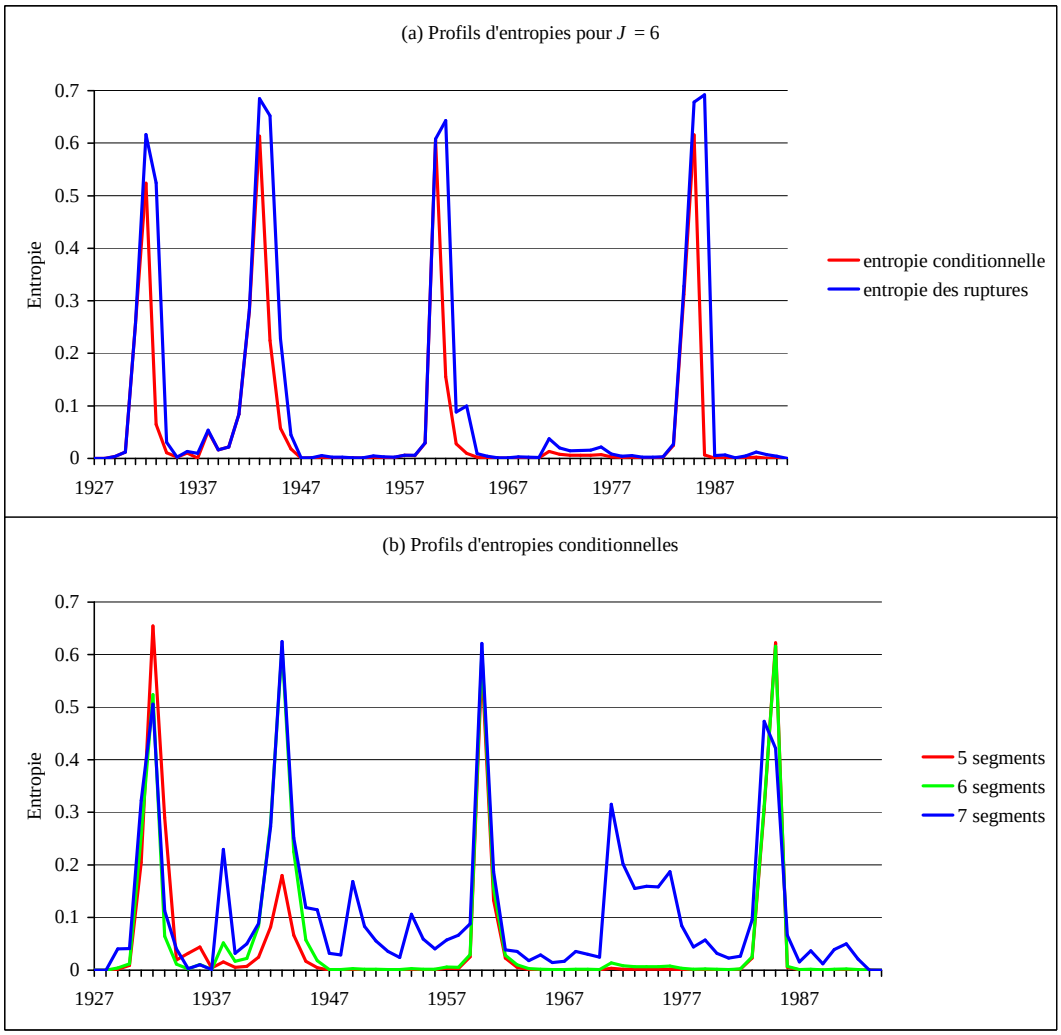

Figure 2. Profils d'entropies.

4 Illustration

L'approche proposée est illustrée sur des données de croissance de pin laricio où la séquence mesurée représente les longueurs de pousses annuelles (segment de tige mis en place sur une année) successives le long du tronc. L'objectif est alors d'identifier les différentes

phases de croissance et d'évaluer la validité de la segmentation ansi obtenue en utilisant un modèle gaussien de changement sur la moyenne et la variance (Guédon *et al.*, 2007).

On remarque que l'écart est grand entre l'entropie des segmentations $H(\mathbf{S}|\mathbf{X}=\mathbf{x};J)$ et l'entropie des ruptures $\sum_{t=1}^{T-1} H(R_t|\mathbf{X}=\mathbf{x};J)$ résumant l'incertitude sur les lois *a posteriori* des ruptures $[\{P(R_t=0|\mathbf{X}=\mathbf{x}), P(R_t=1|\mathbf{X}=\mathbf{x})\}; t=0,\ldots,T-1]$ (Figure 1a). Ceci montre que les lois *a posteriori* des ruptures doivent être interprétées avec précaution dans la mesure où elles ne traduisent que partiellement l'incertitude sur les segmentations du fait de la marginalisation intrinsèque à leur construction et de la dépendance forte dans les modèles de détection de ruptures multiples (indépendance conditionnelle entre le futur et le passé uniquement aux instants de rupture). La divergence de Kullback-Leibler (Figure 1b) montre un changement de pente à $J=6$ ce qui est cohérent avec les profils d'entropies conditionnelles (Figure 2b) où pour $J=7$, l'entropie conditionnelle devient non-négligeable sur de larges plages d'années. La comparaison entre le profil d'entropies conditionnelles et le profil d'entropies des ruptures (Figure 2a) permet de localiser l'écart global entre l'entropie des segmentations et l'entropie des ruptures montré sur la Figure 1a.

Bibliographie

[1] Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics* 86, 221-241.
[2] Cover, T. M. et Thomas, J. A. (2006). *Elements of Information Theory*, 2nd edition. Wiley, Hoboken, NJ.
[3] Fearnhead, P. (2006). Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and Computing* 16(2), 203-213.
[4] Girón, J., Moreno, E. et Casella, G. (2007). Objective Bayesian analysis of multiple changepoints for linear models. In: *Bayesian Statistics* 8, J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith et M. West (Eds.), Oxford University Press, pp. 1-27.
[5] Guédon, Y. (2008). Exploring the segmentation space for the assessment of multiple change-point models. INRIA, Research report RR-6619.
[6] Guédon, Y., Caraglio, Y., Heuret, P., Lebarbier, E. et Meredieu, C. (2007). Analyzing growth components in trees. *Journal of Theoretical Biology* 248(3), 418-447.
[7] Lavielle, M. (1998). Optimal segmentation of random processes. *IEEE Transactions on Signal Processing* 46(5), 1365-1373.
[8] Liu, J. S. et Lawrence, C. E. (1999). Bayesian inference on biopolymer models. *Bioinformatics* 15, 38-52.
[9] Tibshirani, R., Walther, G. et Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society, Series B* 63(2), 411-423.